DOCUMENT RESUME

ED 221 574 TM 820 603

AUTHOR Gersten, Russell
TITLE The Site Variability Issue in Follow Through Revisited: Some New Data and Some Personal Insights.
PUB DATE Mar 82
NOTE 33p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Compensatory Education; *Data Analysis; Demonstration Programs; Federal Programs; *Program Implementation; *Research Design; *Research Methodology; *Research Problems; School Location; Site Analysis
IDENTIFIERS Aggregation (Data); Direct Instruction Follow Through Project; Meta Analysis; Outliers; *Project Follow Through

ABSTRACT

The site variability issue for Direct Instruction Follow Through sites is examined from a variety of perspectives--meta-analysis of the Abt Associates' data, norm-referenced comparisons, and a longitudinal look at stability of effects. Much of the seemingly unexplained variability found in both the Abt report and the reports by M. Kennedy and E. House for Direct Instruction was due to: (1) varying degrees of appropriateness/inappropriateness of the local comparison groups selected, and (2) the inclusion of one site that was not implementing the model. The demographic variables for Follow Through and comparison group indicated two severe mismatches, Tupelo and West Iron County. If two sites and Grand Rapids are eliminated, much of the unexplained variability vanishes. There do appear to be some sites that are more effective than others, yet only one site (Racine) demonstrates weak, negative effects. It appeared that the use data from the pooled comparisons might present a less biased picture--especially in areas such as Reading where home and entry characteristics are fairly heavily weighted. Regression models could be used to derive a predicted or expected score for students in the experimental sample, based on demographic characteristics and pretest information. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED221574

The Site Variability Issue in Follow Through Revisited: Some New Data and Some Personal Insights

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R Gersten

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Russell Gersten

University of Oregon

Paper presented at AERA; New York, NY, March 1982

TM 820 603

The most frequently reached conclusion from the national evaluation of Follow Through undertaken by Abt Associates and Stanford Research Institute (SRI) is some variant of the following statement, "The main lesson of site variability is that the models are not powerful enough to countervail unmeasured, site-specific determinants of outcomes" (Stebbins et al., 1977). For exa mple, House, Glass, McLean, and Walker (1978) used meta-analysis techniques to demonstrate that there was significantly more outcome variance between sites implementing the same educational model (within-model variance) than there was between the various educational models. They went on to conclude that the Follow Through data, therefore, indicate that educational problem-solving and educational change must be conducted primarily at the local level, that no educational model or approach can work everywhere, that each must adapt itself to the needs and exigencies of each individual community.

The major purpose of this paper (and this symposium) is to challenge this assertion by looking carefully at the Abt data on Direct Instruction Follow Through sites, as well as more recent data on these projects. The analyses and perspectives presented owe much to earlier critiques and re-analyses of the Follow Through evaluation--especially those by Becker (1977), Bereiter and Kurland (1981-82), and Kennedy (1978). The vantage point will be a bit different, however.

I began working in Follow Through evaluation in 1979, long after the Abt report and its critiques had been widely discussed and disseminated. My perspective could thus be considered a member of the "second generation" of Follow Through researchers. By the time I began, much of the initial heat and hysteria in the debates of the fairness of the Abt evaluation had died down. A host of classroom-based studies synthesized by Rosenshine (1979) and others seemed to corroborate one of the major findings

of the Abt evaluation--that models that emphasized academic skills and trained teachers to teach these subject areas in a structured, systematic fashion tended to have a positive impact on student achievement in reading and math as measured on standardized achievement tests. Whether the range of measures utilized in the national evaluation was ultimately "fair" to the models espousing a loosely structured or "child-centered" approach had little effect on my day to day activities or the concerns of the school districts with which I worked. The questions I tackled were a bit more mundane than those approached by the Abt group. Rather than attempting to determine which educational model "works best" for all low income children, I began to explore the parameters, and nooks and crannies of the Direct Instruction model. Some of my earlier work looked at the academic progress of children entering Follow Through with low IQs (Gersten, Becker, Heiry & White, Note 1), the later effects of Follow Through (Becker & Gersten, 1982), the relationship between observed level of implementation and academic growth (Gersten, Carnine & Williams, 1982).

I was then asked to prepare reports documenting the effectiveness of each of nine Direct Instruction sites over the last decade for the Joint Dissemination and Review Panel of the Department of Education. To do so, I integrated data from the Abt report (collected in 1973 to 1975) with more recent achievement data collected by the University of Oregon (Becker & Engelmann, 1978) and local testing programs. Since there was no way to continue to collect achievement data from comparison group students over the entire decade, I used a variety of comparison "standards"--occasionally a local comparison school, but often a mean score for low income students in similar schools in that region of the country. I came to know each of the sites intimately, seeing the idiosyncracies of their program components and administrative policies. I was still struck

by the absolute consistency I saw across the nine sites--how the same classroom organization, same curriculum materials, same teaching techniques, same procedures for monitoring student progress, were implemented in sites as diverse as Uvalde, Texas (with a large population of limited-English-speaking students) and Flint, Michigan.

Much of the site variability I observed seemed healthy. I came to see how a strong administrator could raise performance (and thus increase site variability), how antagonism or lack of support from the building principal could decrease performance. I also realized the wisdom of Cook and Campbell's (1979) assertion that no quasi-experimental design will be free of flaws, and that replication is essential. The JDRP reports (Gersten, 1981) clearly demonstrated how the magnitude of treatment effect for a given site could fluctuate depending on the nature of the comparison group (or standard) selected, and the level of achievement test selected for pretest and posttest. But I also saw that, in some sense, time could not be fooled--most of the sites showed some sort of effect over the decade regardless of the comparison standard or type of test used. And invariably, when there were dramatic upward or downward shifts in performance over a two to four year period, the change could be linked to clear problems in (or improvement in) implementation.

After working on such detailed site level analyses, I realized that the problems in aggregating across sites have been severely overlooked (or at least understated) in all the essays and critiques on the Follow Through evaluation.

The question: Is Direct Instruction significantly more effective than other Follow Through approaches? asked by Bereiter and Kurland (1981-82) seemed unanswerable when one merely pooled together data from sites as diverse as Uvalde, Flint, and rural white communities like Flippin, AR? On the other hand, questions such as: Does Direct Instruction have a significant, replicable impact on the math and reading achievement of disadvantaged Black students in urban areas such as Flint, New York, Providence, and East St. Louis? seemed quite reasonable. And the answers are quite clear.

Before looking at the data, I will briefly review some of the major flaws in the Abt/SRI evaluation, and how they contributed to some of the confusions and controversies in interpreting the data.

Flaws in the Original Follow Through Design

With all the perspicacity of hindsight, it is easy to criticize the planned variation design of the national Follow Through experiment. This has been done in such endless detail, that only the highlights will be recounted here. From a technical perspective, the design contained too much variation and too little planning. The performances of nine very different educational models were contrasted across differing samples of children varying on every conceivable demographic variable. The local comparison groups selected by SRI to counteract some of these problems tended to be somewhat mismatched with the Follow Through groups, often less disadvantaged.

Yet it is important to understand that the experiment was designed over a decade ago, in a much more optimistic era. The optimism was reflected both in a naive faith in the ease with which social problems could be eradicated and an equally naive faith, in the power of statistical adjustments to callibrate for initial inequities between treatment and

comparison groups. The latter belief was fueled in large part by the widespread dissemination of Campbell and Stanley's (1966) treatise advocating the use of quasi-experimental designs. Thus, although the assignment (linking) of sites to sponsors was unbalanced and unsystematic by any reasonable experimental design criteria, USOE believed it was far better to be ethically and socially "fair" by linking up sponsors with the communities who were most interested in that approach. Thus, across the 20 competing educational models or sponsors, the number of sites and the nature of the sites varied in a chaotic fashion, making precise comparisons between the models difficult. But the original decision by Egbert to not use random assignment, and rather associate sponsors only with communities showing an interest in implementing the experimental procedures seems to us to have been the only reasonable approach.

On the other hand, the haste and sloppiness with which local comparison groups were selected should be considered a serious design flaw (cf. Haney, 1977). The enormity of these mismatches in some cases is staggering, and one does not get a good feeling for this in the rather polite discussions by Stebbins et al. (1977) or House et al. (1978). For example, in Tupelo, Mississippi, where 95 percent of the Follow Through students were Black and the median family income was $2624, SRI selected a comparison group which was 100 percent white with a median income of $9334. Not surprisingly, results of the Tupelo comparisons were "not significant." It remains unclear why these comparisons were ever made. (There were also cases of quite reasonable comparison groups. In Williamsburg County, South Carolina, for example, both the Follow Through and comparison groups were over 99 percent Black, with family incomes within a few hundred dollars of each other. Entry scores on the WRAT were roughly comparable.)

Abt attempted to deal with the mismatches in a variety of manners. First and foremost was the use of covariance-adjusted means. Yet, Campbell and Erlebacher (1970) demonstrate that covariance adjustments tend to put the poorer children (most often the Follow Through children) at a disadvantage. Aware of this problem, Abt "grayed out" or deleted those cases where there was a serious mismatch, considering the results untrustworthy. Deciding exactly what criteria to use before a comparison is considered untrustworthy was a thorny problem. Stebbins et al. (1977) somewhat arbitrarily selected a 5-point difference in entry performance on the WRAT as the best criterion.

Interestingly enough, the more recent meta-analyses have included all sites, even those with severe "mismatches". As Bereiter and Kurland (Note 1) indicated in 1978, much of the site variability can be explained by variability in the fairness of the local comparison group. The following example demonstrates this rather vividly.

A Closer Look at Site Variability:

The Results in Mathematics

Figure 1a presents the descriptive statistics in MAT Total Math

---

Figure 1a about here

---

for the Entering-Kindergarten (EK) Direct Instruction sites in Cohort III. (West Iron County was deleted since Abt deemed the results untrustworthy; Grand Rapids was deleted since the Direct Instruction model was not implemented.) Median standard scores for each Follow Through and comparison sample have been converted to percentiles. Each box represents one-fourth of a standard deviation unit on the national norm.

Nothing in the Abt report, or the subsequent essays, prepares one for this image of virtually perfect replicability across sites. All the effects were statistically significant; the mean magnitude of effect ranged from .65 to .97 pooled SD units. Note that the comparison (NFT) groups are virtually identical. A look at the demographic data shows that in each community, the comparison students tended to (a) come from families with somewhat higher income, but (b) score lower on the WRAT. (There are two possible reasons for this discrepancy: the WRAT was often administered the sixth to tenth week of school; thus, Follow Through students often had five to nine weeks of instruction before the pretest was given. Also, Follow Through students tended to have a bit more preschool, and hence, in some cases, more exposure to the basic letter and number skills tested on the WRAT.) Yet, the discrepancies on demographic variables do not seem unreasonably high in this case. From Figure 1a, one could conclude that Direct Instruction had a consistent, replicable significant effect on math achievement of urban

Blacks. This pattern was replicated with students from an earlier cohort. Students scored at or above national norm levels, i.e., they reached the objective set by the sponsor.

Figure 1b about here

Figure 1b presents a different picture. Here, all five Direct Instruction sites are included--the previous three urban sites (New York, Flint, MI, and Providence, RI) along with West Iron County, MI, and Grand Rapids, MI. Note how the picture suddenly changes--there is some chaos, clear indications of "site variability."

Note that though West Iron County comes within a couple of points of meeting the objective of performance of the national median, its students appear to perform significantly lower than comparison students. There's a good reason for this. The West Iron comparison group was considered inappropriate by Abt since the NFT students came from higher income families and entered at a significantly higher academic skill level. (In this rural community, students in the school building who were ineligible for Follow Through--usually because their income level was too high--were comprised of the control group.) Abt "grayed out" most West Iron results, since the comparison was unfair. Yet West Iron has been included in the subsequent meta-analyses of Follow Through data (House et al., 1978; Kennedy, 1978)--used to "prove" that program effects are variable.[1] And then there's Grand Rapids...Demographically,

[1] Even if it had been possible to find a suitable comparison group, it is unclear that West Iron County fits in with the other four sites--each of which is a large or middle-sized city with a primarily black population. It seems odd to aggregate this one rural, white site (with a much lower level of poverty) with the others.

Grand Rapids fits nicely in with the three urban, Black sites. But there is a major problem--Grand Rapids was not implementing the Direct Instruction model for most of the students' academic careers. This point has been noted by Kennedy (1978), Becker (1977) and the earlier version of the Abt report (Stebbins, 1976). Documentation exists in the national Follow Through office. And yet Grand Rapids has been included in all analyses and meta-analyses. When Grand Rapids enters the picture, site variability increases. In reality, Grand Rapids would appear to be the exception that proves the "rule"--when the model is not implemented in an urban setting, Follow Through students do not perform significantly different than comparison students.

Depending on the inclusion rules, one sees either a picture of consistent, replicable effects (Figure 1a) or generally positive effects with strong amounts of "unexplained" site variability (Figure 1b). It seems clear in this case that much of the variability can be explained by including a site where the comparison group was inappropriate and including a site that was not implementing the model.

Tables 1 and 2 utilize meta-analysis procedures to examine the same

Tables 1 and 2

variable for all Direct Instruction sites in both cohorts. Following Kennedy's (1978) procedure, the magnitude of the covariance-adjusted treatment effect is expressed in pooled standard deviation units. I have expanded Kennedy's procedure in two ways:

a) utilizing local and pooled comparisons,

b) computing medians as well as means (see Table 2).

The former was done because I felt there would be less chance of a "mismatch" between a given site and the pooled comparison group (con-

sisting of approximately 1000 students in many communities) than between the Follow Through sample and the local comparison group. With the small sample sizes and skewed distributions one encounters when the site is used as the unit of analysis, it also made sense to compute medians as well as means.

Table 1 presents the data for each site; Table 2, descriptive statistics for all EK, all EF and, finally, all sites. Generally, when the magnitude of the effect is .25 pooled SD units or higher, the results are considered educationally significant.

If one begins with the familiar cohort III EK sites (upper right), one sees the same picture--consistent, dramatic, positive effects for the three urban effects, a negative effect for West Iron, and weaker effects for Grand Rapids. (The effects for New York, Flint, and Providence were also statistically significant.) Looking at Cohort II, one does find some variability. Flint and East St. Louis are positive for both local and pooled comparisons. New York shows a strong effect when compared to the pooled comparison group on the basis of demographic characteristics, but a weak effect when compared to the local comparison group. The local comparison was grayed out by Abt; median percentile for New York students was the 60th, well above grade level. Note that the effects for the other urban sites do not fluctuate from local to pooled comparisons. Racine shows slight negative effects; it seems reasonable to conclude that the Racine site did not show the same positive effects as the other Direct Instruction projects. (Grand Rapids' fluctuations are of little interest, since it is unclear what is being measured.) Thus, with the exception of one cohort in Racine,

whether one uses adjusted or unadjusted scores, one gets a picture of consistent, strong replicable effects in Total Math for the EK programs. The magnitude of effect is .40 pooled SD units (Table 2). Most of the variability can be explained by issues in sampling and inclusion of a site not implementing the program. It would seem fair to say that students in four-year Direct Instruction programs in urban sites performed at the national median level, and (with two exceptions) significantly above local comparison groups. With one exception (Racine), they perform at a significantly higher level than one would predict low income students to perform based on their demographic characteristics (i.e., when compared to the "pooled comparison groups").

I will deal with the EF sites more briefly. Tupelo is the serious mismathc here. It has been discussed earlier in this paper. Note that scores rise in Tupelo when the pooled comparison group is used. For a further discussion of Tupelo, see Becker (Note 1). Otherwise, the program appears to show consistent strong effects.

A Look at Reading Comprehension

The next section will examine the results on the Reading (Comprehension) subtest of the MAT from a variety of perspectives. This subtest was selected for several reasons. It was one of the five subtests on which Kennedy (1978) performed meta-analyses, and I wish to use her work as a basis for expansion. In addition, many teachers and administrators have perceived the Reading Comprehension program as one of the strengths in the Direct Instruction programs, and one of the areas in which they have seen greatest growth. (The same is not true for Word Knowledge, Vocabulary, or Spelling.) The data from both the Abt analyses and the Kennedy meta-analyses are inconclusive in this domain. Kennedy reports a mean magnitude of effect of .07 SD units. The Abt data indi-

cates that though Direct Instruction is the first ranked of Follow Through sponsors, its students significantly outperformed comparison students only 31% of the time (Becker, 1977).

A perusal of the regression weights used in the Abt report reveals that the coefficients were often triple the value for Reading Comprehenson than for other subtests (e.g., the Math and the Self-Concept measures). Thus, the bias due to regression artifacts in statistical tests involving the local comparison samples were likely to be most extreme in this domain.

In discussing the Reading Comprehension data, I wish to develop the idea of the use of "expected" (or predicted) scores in evaluation research. Wtih this approach, a predicted score is derived for each sample based on demographic characteristics. A multiple regression model is used with a large sample of comparison children who come from communities globally similar to the local site in question. The advantage of this approach is that effects of a local "mismatch" will be minimized or eradicated, because of the size of the sample. (While any one or two local comparison schools may be inappropriate, as was the case in Tupelo, or may have had an exemplary program, as was the case in New York City, it is highly unlikely that this will happen with all the schools in the large pooled sample.

This approach was utilized by Stebbins et al. (1976, 1977) to compensate for the irregularities and imperfection in some of the local comparison groups. Each Follow Through sample was compared to a "pooled comparison group" (i.e., all NFT children in all sites) using covariance procedures. In effect, each Follow Through sample

was compared to a score based on all non-Follow Through samples with somewhat similar demographic characteristics.

My contention is that the pooled comparisons often present a more accurate picture of what happened--particularly in domains such as Reading Comprehension where there was a large potential for problems due to regression artifacts. Table 3 demonstrates this.

Table 3 presents the percent of statistically and educationally significant positive effects for each sponsor--the left column uses local comparisons, the right column the pooled comparison group. When compared to local groups, Direct Instruction appears to be neutral, paralleling Kennedy's finding of a mean effect of .07; yet, when the pooled group is used, Direct Instruction produces significant positive effects in 47% of the sites. (Nothing this dramatic happens in Math.)

Table 4 demonstrates effects for each site using the covariance-adjusted pooled comparisons. Note that the magnitudes are generally not as strong as in Total Math, but one sees the same basic patterns. With the exception of Racine, all EK sites show positive effects. The effect in Providence is too weak to be considered educationally significant. As expected, Grand Rapids performs at about the same level as the comparison group. Except for Williamsburg, the magnitude of effects are not as strong in Reading as Math for the EF sites; but again, they are consistently positive.

I do not believe the data support Kennedy's assertion that Direct Instruction programs were more effective in the three year programs. I believe this should be rephrased. "Meta-analyses show Williamsburg County was the most successful Direct Instruction site."

In my opinion, the use of pooled comparison groups, or "expected" scores, is currently underused in evaluation research. There will always be a problem in finding a fair, appropriate local control school. Even if one avoids the absurdities of the Tupelo situation, there is always a chance of selecting a comparison school that is implementing a strong, innovative program, rather than "representing existing, traditional practices." There also is the possibility of selecting a comparison school that is less effective than most other schools in the district. By using "pooled groups" such as all Title I eligible schools in Flint or Providence or rural South Carolina, and deriving a predicted score for a Follow Through sample using demographic information, one can get a good counterbalance. Furthermore, this information is becoming increasingly more available. The failure of the metaanalyses by House et al. and Kennedy to include this data does not make sense.

Norm Referenced Comparisons: The Stability of Effects

Table 5 demonstrates the norm-referenced data collected by Abt on cohorts II and III in Total Reading. The median percentile for Direct Instruction sites is the 41st for cohort II and 40th for cohort III. This is a fourth of a standard deviation unit below the national norm level; it is over half a standard deviation unit above the median for the other sponsors--the 23rd percentile. Studies of low income schools serving primarily minority students show that third grade reading performance is typically at the 22nd to 28th percentile level (Ozenne et al., 1976). This is the level at which Gradn Rapids performs. Thus, the results are not as dramatic as in Math; there is evidence that all the sites, with the exception of Grand Rapids, are above the expected level. Two sites are somewhat below the 41st percentile--Racine and Providence, and one site is appreciably above (New York City). But note that the magnitudes of these fluctuations are within one-fourth SD unit. The only exception is the superlative performance of cohort II in New York. This type of "unexplained" variability--above the national norm level--is hardly a problem.

Table 6 demonstrates the progress of the six urban Direct Instruction EK sites who have remained with the model since 1974. Note that the cast of characters changes somewhat. Grand Rapids, Racine, Providence, and West Iron County are gone (since they ended their affiliation with Direct Instruction in the 1974-77 period), and one site not included in the Abt evaluation, Washington, D.C., is added. There are a few other differences between the data on this table and the Abt data:

(a) only low income students are included in our analyses whereas Abt

included some non-low income students,

(b) cohorts are numbered differently--our cohort II is Abt's cohort I, etc.,

(c) mean rather than median standard scores are used in our analyses,

(d) up until 1977, the MAT was used at all sites. From 1978-81, local test data was utilized. Those later cohorts were tested on a wide range of instruments--CAT, CTBS and SRA.

The most striking aspect of these data is that the performance of all four urban sites--New York, Flint, East St. Louis, and Washington, D.C.--tends to center around the 40th percentile. The basic performance level reached in the years of the Abt evaluation is maintained. There are individual site fluctuations. In East St. Louis, one notes a steady increase. Washington D.C. tends to be consistently a low performer. This is borne out by information on level of implementation at that site. New York demonstrates a decline in the 1978 to 1979 period, followed by a rise in the last two years. This apparently was due to some of the chaos related to the New York City budget crisis and other administrative problems which have since been resolved. . The major point is that most of the variability can be explained and understood. These phenomena will recur in any implementation of a model in the field. The magnitude of these variations is rarely over one-fourth SD unit. When it is larger, it is often in a positive direction.

Table 7 displays the data for Total Math. New York was omitted because the city did not test math. Here, the scores fluctuate around the 50th percentile, the national median. All scores are within one-

fourth deviation of this point except for 1981 in East St. Louis, which is appreciably higher. Note how the norm-referenced data parallels the findings from the meta-analysis; that is, the level of performance is consistently somewhat higher in Math than Reading. Again, there appears to be very little unexplained variability.

Summary and Conclusions

This paper looked at the site variability issue for Direct Instruction Follow Through sites from a variety of perspectives--meta-analysis of the Abt data, norm-referenced comparisons, and a longitudinal look at stability of effects. With a careful, multi-perspective approach, several things became clear. The first is that, as Becker (1977) and Bereiter and Kurland (1981-82) indicated, much of the seemingly unexplained variability found in both the Abt report and the Kennedy and House reports for Direct Instruction was due to

(a) varying degrees of appropriateness/inappropriateness of the local comparison groups selected by SRI,

(b) the inclusion of one site that was not implementing the model. This is particularly clear when one looks at the actual norm-referenced achievement data collected by Abt and the University of Oregon. Performance in Total Math for low income students in four-year Follow Through programs fluctuates within one-fourth standard deviation unit of the national median; performance in Reading fluctuates around the 40th percentile. There are few exceptions, and more of these are in a positive than a negative direction. Data were also looked at using the meta-analysis procedures used by Kennedy. However, effect sizes were estimated using both local and pooled comparison groups. A look at the demographic variables for Follow Through

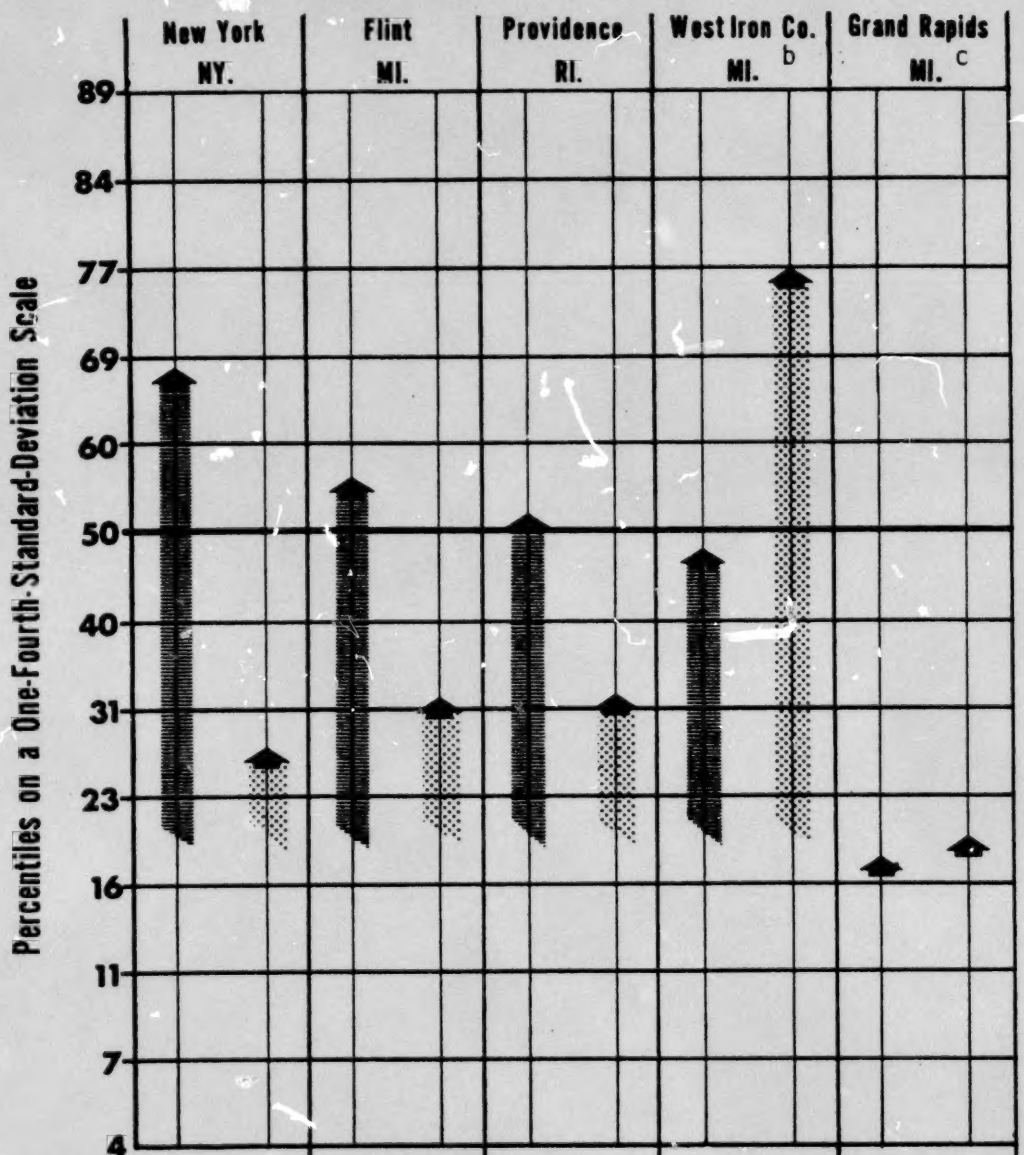
and comparison group indicated two severe mismatches, Tupelo and West Iron County. In each case, the income level of the comparison group and WRAT entry scores were significantly higher. Each site demonstrated large negative effects in the meta-analysis; this was quite likely due to problems in utilizing statistical models to adjust for samples with widely different demographic characteristics. If two sites and Grand Rapids are eliminated, much of the unexplained variability vanishes. There do appear to be some sites that are more effective than others, however. Yet only one site (Racine) demonstrates weak, negative effects.

It appeared that the use data from the pooled comparisons might present a less biased picture--especially in areas such as Reading where home and entry characteristics are fairly heavily weighted. It would seem wise for future evaluation efforts of the pooled comparison group idea...though in a more refined way than Abt did. Rather than select one local comparison school for a compensatory education program, it might be fairere (and perhaps easier) to use a large sample of Title I eligible schools in that city or region of the country. Regression models could be used to derive a predicted or expected score for students in the experimental sample, based on demographic characteristics and pretest information. This may reduce the problem of matching an experimental program with a school that is dissimilar in some important demographic consideration or a school with a particularly strong or weak program. In some small, rural communities, it may be impossible to find a suitable comparison group. In these cases, it would be wiser to not choose a comparison group at all, rather than pick an inappropriate one.

Finally, future evaluation research should exercise more care in how data are aggregated. While it makes sense to look at the effects of Direct Instruction in seven urban sites with primarily minority students, it seems unwise to throw one rural, Anglo site "into the same pot." Data from the eight urban sites can answer a meaningful question; it is unclear what new information is added from the rural site. And the inclusion of a site where students enter at a much higher level is likely to add to the "unexplained variability." Kennedy's (1978) assertion that effects of each model are outcome specific is borne out by this data. Mean effects are appreciably higher in Total Math than Reading (Comprehension). (They are even higher in Language, and even lower in Word Knowledge, Vocabulary.) Implications of these findings are discussed by Becker (1977). Yet, even in Reading Comprehension, the mean magnitude of effect across all sites is .25 SD units. In Total Math, it is well over one-half of a standard deviation unit. As Resnick (Note 1) indicated, perhaps the most surprising thing about the Follow Through evaluation was that despite the technical and methodological errors and controversies, the results came out so clearly.

Figure 1b

Median Percentile Rankings for Total Math
for all Direct Instruction sites, Cohort III[a]



N	FT	Comp	FT	Comp	FT	Comp	FT	Comp	FT	Comp
%-tile	68th	28th	56th	32nd	52nd	32nd	48th	77th	18th	20th
S. S.	75.8	64.1	72.3	65.0	71	65	70.3	80.2	60.3	61.2

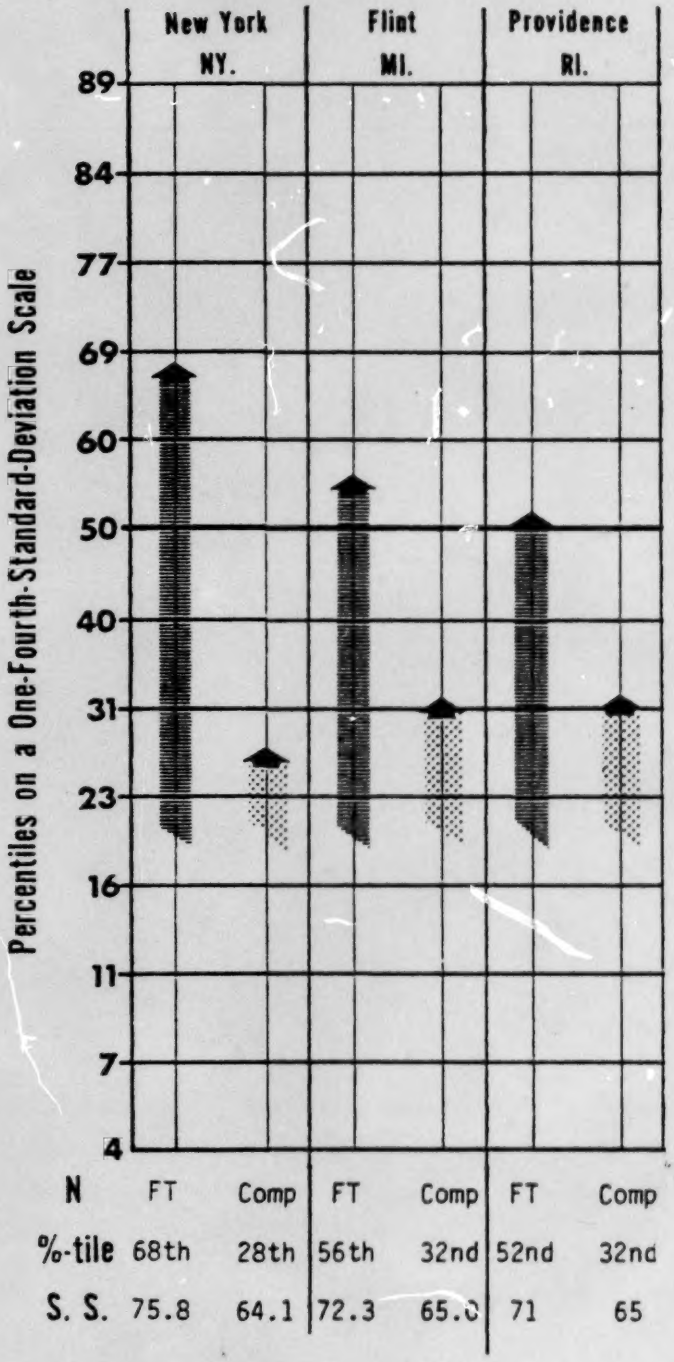
[a]Adapted from Stebbins et al., 1977, 4A, p. 276

[b]Grey out by Abt because of unfair comparison group

[c]Not implementing Direct Instruction model

Figure 1a

Median Percentile Rankings for Total Math, Direct Instruction EK sites, Cohort III[a]



[a] Adapted from Stebbins et al., 1977, Vol. 4A, p. 276.

Table 1

Magnitude of Effects in Pooled SD Units

for MAT Total Math

for all Direct Instruction Sites

Entering-Kindergarten Sites

	Cohort II			Cohort III		
	Local		Pooled	Local		Pooled
New York City	.09	→	.58	.82	→	1.09
Flint	.88		.82	.65		.71
E. St. Louis	1.24		1.20	.97		.97
Racine	-.18		-.13	-.52[a]		-.53
Grand Rapids[b]	.39	←	-.29	.30		.33

Entering First Grade Sites

	Local		Pooled	Local		Pooled
Williamsburg County	.42	→	1.58	1.22		1.13
E. St. Louis	1.08	→	1.31	.50		.53
Tupelo[a]	-1.28	→	-.40	-.39	→	.16

→ indicates the magnitude of the effect increases substantially more when pooled comparison group used

magnitude becomes appreciably less when pooled comparison group used

[a] Case of clear mismatch

[b] Not implementing Direct Instruction Model

Table 2

Mean and Median Magnitude of Effects

in MAT Total Math for Direct Instruction Sites

		LOCAL		POOLED	
	N	Mean	Median	Pooled Mean	Median
Entering-K sites	10	.40	.52	.41	.65
Entering-First sites	6	.26	.46	.72	.83
All Sites	16	.35	.58	.53	.64

Table 3

Percent of Significant Outcomes on

MAT Reading (Comprehension)-Elementary Level

	Comparisons with Local Control (NFT) Group	Comparisons with Expected Score[a]
Direct Instruction	0	+47
Responsive Education	0	+10
Parent Education	-9	+10
Bank Street	0	+10
SEDL	+17	0
Cognitive Curriculum	-40	0
Behavior Analysis	-17	-22
TEEM	-22	-33
EDC	-25	-43

[a] Expected score based on general linear regression model using WRAT entry, SES, ethnic group, and parent education as predictors and using the sample of 6000 Non-FT children as a comparison sample

Table 4

Magnitude of Effects in Pooled SD Units for MAT Reading (Comp.)

for all Direct Instruction Sites

Entering-Kindergarten Sites - Pooled Comparisons Only

Cohort II		Cohort III	
New York City	.56	New York City	.25
Flint	+.40	Flint	+.32
E. St. Louis	.44	Providence	.20
Racine	-.20	West Iron[a]	-.58
Grand Rapids[b]	0	Grand Rapids[b]	.13

Entering First Grade Sites

Williamsburg County	1.81	Williamsburg Co.	.76
E. St. Louis	.22	Dayton	.24
Tupelo[a]	-.20	Tupelo[a]	0

Summary Statistics

	N	M	Mdn
All EK sites	10	.15	.23
All EF sites	6	.50	.23
All Sites	16	.28	.25

[a] Case of clear mismatch

[b] Not implementing Direct Instruction Model

Table 5

Total Reading Scores for All Four Year

Follow Through Sites (from Stebbins et al., 1977)

	Cohort II		Cohort III	
Site	Median %ile	% of Students 1 or more yrs Below Grade Level	Median %ile	% of Students 1 or more yrs Below Grade Level
New York, NY	58th	9.7	48th	16.1
E. St. Louis, IL	45th	17.9	b	b
Racine, WI	37th	21.0	b	b
Flint, MI	41st	16.9	42nd	7.1
Grand Rapids, MI[a]	28th	39.8	28th	40.7
Providence, RI	b	b	34th	25.3
West Iron Co., MI	b	b	48th	11.8
All DI Sites	41st	--	40th	20.2
All FT Sites	23rd	--	--	--
MAT Norm Sample	50th	20	50th	20

[a] Not implementing the DI Model

[b] Cohort not included in the Evaluation

Table 6

Stability of Effects: Percentile Equivalents[1] for

End of Third Grade Total Reading Scores

for Urban Black Four-Year Direct Instruction Follow Through Projects[2]

Site	Cohort Finished In:	II 1973	III 74	IV 75	V 76	VI 77	VII 78	VIII 79	IX 80	X 81
E. St. Louis IL		--	42	35	41	36	--	54[a]	55[a]	66
Flint, MI		34	35	40	40	34	46	53	52	42[d]
New York, NY		36	52	42	46	40	33[b]	36[b]	46[b]	47
Washington, DC		34	30	40	34	30	44[c]	44[c]	--	--

[a] CAT

[b] SAT

[c] CTBS

[d] SRA

Note: The mean score for non Follow Through low-income students in large Northeast cities corresponds to the 28th percentile on the Metropolitan

1. Mean Standard Score converted to percentile

2. The MAT (1970 Elementary) was given at all sites for cohorts 3 to 6 (1974 to 1977). In the 1978-80 period, a variety of standardized tests were administered, depending on local policy. The Anchor Tables (Loret et al., 1979) and some of our own research established the comparability (within 2 to 3 percentile points).

Table 7

Stability of Effects: Percentile Equivalents for End of Third Grade Total Math Scores for Urban Black Direct Instruction Sites with Kindergarten-to-Third Grade Programs

Site	Finished 3rd grade in: 1973	74	75	76	77	78	79	80	81
E. St. Louis IL	56	57	62	61	50	--	52[a]	59[a]	72[a]
Flint, MI	--	46	52	52	57	57[b]	55[b]	61[b]	53[b]
Washington, DC	40	52	66	52	52	56	53[c]	60[c]	--

[a] CAT

[b] SRA

[c] CTBS

Note: The mean score for non-Follow Through low-income students in large Northeast cities corresponds to the 18th percentile on the Metropolitan

Reference Notes

Becker, W. C. Research on direct instruction, book in preparation.

Gersten, R., Becker, W. C., Heiry, T. J., and White, W. A. Relationship between entry IQ and yearly academic gains in a direct instruction model. Paper presented at AERA; Boston, Mass., April 1981.

Resnick, L. Personal communication, 1979.

References

Becker, W. C. The national evaluation of Follow Through--Behavior-theory-based programs come out on top. Education and Urban Society, 1978.

Becker, W. C., & Engelmann, S. Analysis of achievement data on six cohorts of low income children from 20 school districts in the University of Oregon Direct Instruction Follow Through Model. Follow Through Project, Technical Report #78-1. Eugene, OR: University of Oregon, 1978.

Becker, W., & Gersten, R. A follow-up of Follow Through: The later effects of the direct instruction model on children in fifth and sixth grades. American Education Research Journal, in press.

Bereiter, C., & Kurland, M. A constructive look at Follow Through results. Interchange, 1981-82, 12, 1-22.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1963.

Cook, T. D., & Campbell, D. T. Quasi-experimentation: Design and analysis for field settings. Chicago: Rand McNally College Publishing Company, 1979.

Gersten, R., et al. Final reports to Joint Dissemination Review Panel, National Institute of Education, for Direct Instruction Follow Through Projects in Williamsburg Co., SC; DeKalb Co., TN; Flippin, AR; P. S. 137, NY; Cherokee, NC; Uvalde, TX; E. Las Vegas, NM; Nichols Avenue School, Washington, D.C. In Educational programs that work: A catalog of exemplary programs approved by the joint dissemination review panel. (8th ed.) Department of Education, National Diffusion Network Division, 1981.

Gersten, R. M., Carnine, D. W. & Williams, P. B. Measuring implementation of a structured educational model in an urban school district: An observational approach, Educational Evaluation and Policy Analysis, 1982.

House, E., Glass, G., McLean, L., & Walker, D. No simple answer: Critique of the Follow Through evaluation. Harvard Educational Review, 1978, 48(2), 128-160.

Kennedy, M. Findings from the Follow Through planned variation study. Educational Researcher, 1978, 1, 3-11.

Ozenne, D., et al. United States Office of Education. Annual evaluation report on programs administered by the U.S. Office of Education, FY 1975. Washington, D.C.: Capital Publications, Educational Resources Division, 1976.

Rosenshine, B. V. Content, time, and Direct Instruction. In Penelope L. Peterson and Herbert Walberg (Eds.), Research on teaching: Concepts, findings, and implications. Berkeley: McCutchan Publishing Corp., 1979.

Stebbins, L. B. (ed.) Education as experimentation: A planned variation model. Vol. IIIA. Cambridge, MA: Abt Associates, Inc., 1976.

Stebbins, L. B., St. Pierre, R. G., Proper, E. C., Anderson, R. B., & Cerva, T. R. Education as experimentation: A planned variation model. Vols. IV-A-D. An evaluation of Follow Through. Cambridge, MA: Abt Associates, 1977.